USE OF PARTICLE FILTERING AND MCMC FOR INFERENCE IN PROBABILISTIC ACOUSTIC TUBE MODEL

Ruobai Wang* *Yang Zhang*† *Zhijian Ou** *Mark Hasegawa-Johnson*†

* Tsinghua University, Department of Electronic Engineering
† University of Illinois, Urbana-Champaign, Department of Electrical and Computer Engineering
wangruobai11@tsinghua.org.cn, yzhan143@illinois.edu, ozj@tsinghua.edu.cn, jhasegaw@illinois.edu

ABSTRACT

The Probabilistic Acoustic Tube (PAT) model is a probabilistic generative model of speech. By associating every generative parameter with a probability distribution, it becomes possible to convert every standard speech analysis task into a probabilistic inference task, thereby grounding every such task with quantifiable measures of bias and consistency. The previously published PAT model did not adequately model AM-FM and therefore phase of the voice source. In this paper, we model the AM-FM of the voice source using an autoregressive process. The resulting model is a non-linear state-space model and thus has no closed-form inference algorithm, but effective inference can be achieved by using Auxiliary Particle Filtering (APF) and Taylor expansion assisted Markov Chain Monte Carlo (MCMC). Results demonstrate that, unlike previous speech models, this model is able to account for the phase of the voice source, achieving signal reconstruction with 8.79dB SNR.

Index Terms— Speech modeling, particle filter, MCMC

1. INTRODUCTION

In speech analysis and synthesis, a complete model, which jointly considers all speech parameters, would be more useful than a partial model. For example, pitch and spectral envelope [1], vocal tract and glottal source [2], should be considered together. Degottex et. al. [3] proposed a speech model called SVLN, using pitch, glottal source and vocal tract together as its parameters. Turner et. al. [4] proposed the PAFD method, which separates the voiced signal into AM-FM sinusoidal components. The STRAIGHT model [5] also considers pitch, spectral envelope, and glottal source jointly. Most of these efforts, however, estimate parameters separately. A speech model that is capable of jointly inferring all speech parameters in a structural manner is expected to yield higher consistency and accuracy.

Therefore, we have proposed a probabilistic generative model for speech called Probabilistic Acoustic Tube (PAT) [6, 7]. It jointly models breathiness, pitch, glottal excitation and vocal tract, notably with phase information. In PAT3 [8], we incorporated AM-FM effects using Gaussian approximation similar to Bayesian Spectral Estimation (BSE) [9]. The Gaussian approximation, however, is inaccurate especially when the instant amplitude is small [4]. Therefore, in this paper, we model the AM-FM of the voice source using an autoregressive (AR) process. Based on Auxiliary Particle Filtering (APF) and Markov Chain Monte Carlo (MCMC), we successfully develop an effective inference algorithm for this improved but complicated model.

The inference scheme of the new PAT3 consists of two layers. The outer layer applies APF to infer the frame-level hidden variables, such as pitch, group delay, glottal vibration, vocal tract transfer function, and voiced/unvoiced amplitude. The inner layer uses Taylor expansion assisted MCMC to infer the sample-level hidden variables, which are fine voiced amplitude and pitch variations affected by AM-FM effect [13, 14], instead of making Gaussian approximation as in the previous PAT3 model.

The rest of the paper is arranged as follows: Section 2 briefly introduces the model of PAT3. Sections 3 describes the inner and outer layers of the inference scheme respectively. Section 4 shows the experiment results which demonstrate the effectiveness of the new PAT3. Section 5 gives the conclusion and future research directions.

2. MODEL OF PAT3

2.1. The Signal Model

The signal model of the new PAT3 is similar to the original one [8]. It is a source-filter model, where the source is a mixture of glottal vibration and breathy noise, and the filter is the vocal tract. A discrete-time speech frame $x(t)$ is modeled as

$$x(t) = \Big[(v(t) + \tilde{b}u(t)) * h(t) + \tilde{e}\varepsilon(t)\Big]\, w(t) \tag{1}$$

where $*$ represents convolution, $v(t)$ is the voiced source, $\tilde{b}$ is the unvoiced amplitude, and $\tilde{e}$ is the noise amplitude. $h(t)$ is the impulse response of the vocal tract. Both the unvoiced signal $u(t)$ and the noise $\varepsilon(t)$ are modeled by standard white Gaussian processes. $w(t)$ is the rectangular window function.

The voiced source, $v(t)$, is modeled as a superposition of quasi-periodic harmonics affected by AM-FM:

$$v(t) = \tilde{a}(t) \sum_{d=1}^{D} \text{real}\left[G(d\omega_0)\exp(-d\omega_0\tau + d\phi(t))\right] \tag{2}$$

where d is the harmonic order, and D is its maximum. τ is the group delay, and $\omega_0 = 2\pi f_0/F_s$ is the fundamental angular frequency. F_s is the sampling rate. $\phi(t) = \sum_{s=1}^{t} \frac{2\pi}{F_s} f(s)$ is the instant phase of the fundamental harmonic, which is derived from the instant fundamental frequency $f(s)$. s runs through sample points. $\tilde{a}(t) = \log\left(1 + e^{a(t)}\right)$ is the non-negative AM envelope of the voiced source, which is determined from the transformed-modulator $a(t)$ by [4]. $G(\omega)$ is the glottal transfer function. We use the same three-pole model as in [8], and denote its parameters as $\boldsymbol{g}$.

The vocal tract transfer function $H(\omega)$, which is assumed to be causal, can be represented by its real cepstral coefficients $\boldsymbol{h}$ at positive low-quefrency [15].

This project is supported by NSFC grant 61473168 and AHRQ grant R21-HS022948. Corresponding author: Z.O.

2.2. The Probabilistic Model

From the signal model, PAT3 further builds a probabilistic generative model of speech, which essentially is a non-linear state-space model.

2.2.1. *Observation Likelihood*

Transforming Eq. (1) into frequency domain by length-L FFT, and adding subscripts n to index frames, we obtain

$$X_n(\omega) = \Big[(V_n(\omega) + \tilde{b}_n U_n(\omega))H_n(\omega) + \tilde{e}_n E_n(\omega)\Big] \circledast W(\omega), \tag{3}$$

where $\circledast$ represents circular convolution. Define the voiced part $S_n(\omega)$ as

$$S_n(\omega) = (V_n(\omega)H_n(\omega)) \circledast W(\omega) \tag{4}$$

As detailed in [7], we stack the real parts and then the non-zero imaginary parts of $X_n(\omega)$, $\omega = \frac{2\pi}{L}k, k = 0, \ldots, \frac{L}{2}$, into a length-$L$ vector $\boldsymbol{x}_n$, and similarly $\boldsymbol{s}_n$ for $S_n(\omega)$. It can be proved that elements of $\boldsymbol{x}_n$ given $\boldsymbol{s}_n$ are independent Gaussians

$$\log p(\boldsymbol{x}_n|\boldsymbol{s}_n) = \sum_{k=0}^{L-1}\left[-\frac{1}{2}\log(2\pi\sigma_{n,k}) - \frac{\left|X_n(\frac{2\pi}{L}k) - S_n(\frac{2\pi}{L}k)\right|^2}{2\sigma^2_{n,k}}\right] \tag{5}$$

where

$$\sigma^2_{n,k} = \tilde{b}^2_n\left|H_n\left(\frac{2\pi}{L}k\right)\right|^2 + \tilde{e}^2_n \tag{6}$$

2.2.2. *State Transition*

There are two sets of states (i.e. hidden variables). The first set,

$$\boldsymbol{O}_n = \{f_{0,n}, \tau_n, \boldsymbol{h}_n, \boldsymbol{g}_n, b_n\} \tag{7}$$

where $b_n = \log \tilde{b}_n$, is the set of *frame-level* hidden variables, because they are constant within a frame. The letter 'O' stands for 'Outer' for reasons that will be clear soon. Note that $\tilde{e}^2_n$ are estimated directly from the first several frames of an utterance that are assumed to be speech-free, and thus not treated as hidden. The second set,

$$\boldsymbol{I}_n = \{\boldsymbol{I}_n(t)\}_{t=1:T} = \left\{[a_n(t), f_n(t)]^T\right\}_{t=1:T} \tag{8}$$

where the letter 'I' stands for 'Inner', is the set of *sample-level* hidden variables (indexed by t) within a frame. $a : b$ denotes a set of consecutive integers from a to b.

For the first set, we assume Brownian motion on frame level [1]

$$p(\boldsymbol{O}_n|\boldsymbol{O}_{n-1}) = \mathcal{N}\left(\boldsymbol{O}_{n-1}, \Sigma_O\right) \tag{9}$$

For the second set, we apply 2nd-order AR assumption on sample level, inspired by the work in [4] for AM-FM demodulation

$$\begin{aligned} &p(\boldsymbol{I}_n(t)|\boldsymbol{I}_n(t-2:t-1)) \\ =&\mathcal{N}\left(\lambda\boldsymbol{I}_n(t-1) + (1-\lambda)\boldsymbol{I}_n(t-2), \Sigma_I\right) \end{aligned} \tag{10}$$

Here Σ_O, Σ_I are diagonal covariance matrices. For both sets, we assume uninformative priors for boundary conditions.

To sum up, notice that $\boldsymbol{s}_n$ is determined by $\boldsymbol{O}_n$ and $\boldsymbol{I}_n$, the likelihood in Eq. (5) essentially specifies $p(\boldsymbol{x}_n|\boldsymbol{O}_n, \boldsymbol{I}_n)$. Then Eqs. (5), (9) and (10) define the probabilistic model of PAT3 as a non-linear state-space model.

3. DUAL-LAYER MONTE CARLO INFERENCE

A rationale in PAT modeling is that the observed speech can be analyzed by performing inference over hidden variables. Here we are interested in MAP inference, i.e., finding the MAP sequence

$$\hat{\boldsymbol{O}}^{\text{MAP}}_{1:N} \triangleq \underset{\boldsymbol{O}_{1:N}}{\text{argmax}}\, p(\boldsymbol{O}_{1:N}|\boldsymbol{x}_{1:N}) \tag{11}$$

Eq. (11) has no analytic solution for the aforementioned nonlinear state-space model, therefore we adopt the method developed in [16], which is a sequential Monte Carlo approximation to Eq. (11). The general idea is to first perform particle filtering and view the set of resulting particles as a discretization of the state space, and then apply Viterbi decoding.

In order to obtain the filtering distribution $p(\boldsymbol{O}_{1:N}|\boldsymbol{x}_{1:N})$, we need to evaluate the likelihood $p(\boldsymbol{x}_n|\boldsymbol{O}_n)$, which is estimated by Viterbi approximation:

$$p(\boldsymbol{x}_n|\boldsymbol{O}_n) \approx \max_{\boldsymbol{I}_n} \boldsymbol{p}(\boldsymbol{x}_n|\boldsymbol{O}_n, \boldsymbol{I}_n)p(\boldsymbol{I}_n) = \boldsymbol{p}(\boldsymbol{x}_n|\boldsymbol{O}_n, \hat{\boldsymbol{I}}_n)p(\hat{\boldsymbol{I}}_n) \tag{12}$$

where $\hat{\boldsymbol{I}}_n$ is picked from the samples $\left\{\boldsymbol{I}^{(r)}_n\right\}_{r=1:R}$ drawn from $p(\boldsymbol{I}_n|\boldsymbol{O}_n, \boldsymbol{x}_n)$. This amounts to (sample-level) smoothing inference. We apply different Monte Carlo techniques to solve the (sampling-level) smoothing and (frame-level) filtering inference problems, which can be divided into two layers and detailed in the next two subsections respectively.

Inner Layer: Estimate $p(\boldsymbol{I}_n|\boldsymbol{O}_n, \boldsymbol{x}_n)$ using Markov Chain Monte Carlo (MCMC);

Outer Layer: Estimate $p(\boldsymbol{O}_{1:N}|\boldsymbol{x}_{1:N})$ using Auxiliary Particle Filter (APF).

3.1. Outer Layer Inference using APF

Suppose that $p(\boldsymbol{O}_{n-1}|\boldsymbol{x}_{1:n-1})$ is approximated by a discrete distribution:

$$\left\{\boldsymbol{O}^{(i)}_{n-1}, w^{(i)}_{n-1}\right\}_{i=1:M} \tag{13}$$

where $\boldsymbol{O}^{(i)}_{n-1}$ is the i-th sample of $\boldsymbol{O}_{n-1}$ and $w^{(i)}_{n-1}$ is its weight. APF aims to sample a mixture distribution

$$p(\boldsymbol{O}_n|\boldsymbol{x}_{1:n}) \propto p(\boldsymbol{x}_n|\boldsymbol{O}_n)\sum_{i=1}^{M} p(\boldsymbol{O}_n|\boldsymbol{O}^{(i)}_{n-1})w^{(i)}_{n-1} \tag{14}$$

with proposal

$$q(\boldsymbol{O}_n|\boldsymbol{O}^{(i)}_{n-1}, \boldsymbol{x}_n)\beta^{(i)}_n \tag{15}$$

The APF procedure consists of iterating the following steps ($n = 1, \ldots, N$):

1. Resample: Sample M indices $j_1, \ldots, j_M$ from $\{1, \ldots, M\}$ according to $\left\{\beta^{(i)}_n\right\}_{i=1:M}$.

2. Propagate: Sample $\boldsymbol{O}^{(i)}_n$ from $q(\boldsymbol{O}_n|\boldsymbol{O}^{(j_i)}_{n-1}, \boldsymbol{x}_n)$.

3. Reweight: Assign each particle $\boldsymbol{O}^{(i)}_n$ the corresponding importance weight:

$$w^{(i)}_n \propto \frac{p(\boldsymbol{x}_n|\boldsymbol{O}^{(i)}_n)p(\boldsymbol{O}^{(i)}_n|\boldsymbol{O}^{(j_i)}_{n-1})w^{(j_i)}_{n-1}}{q(\boldsymbol{O}^{(i)}_n|\boldsymbol{O}^{(j_i)}_{n-1}, \boldsymbol{x}_n)\beta^{(j_i)}_n} \tag{16}$$

The efficiency of an APF scheme rests primarily on the choice of proposal distribution $q(\boldsymbol{O}_n|\boldsymbol{O}^{(i)}_{n-1}, \boldsymbol{x}_n)$ and resampling weights $\beta^{(i)}_n$. We follow the common practice to set $\beta^{(i)}_n$ proportional to $p(\boldsymbol{x}_n|\boldsymbol{O}^{(i)}_{n-1})w^{(i)}_{n-1}$, and carefully design different proposal distributions for different hidden variables.

[1] Except for τ_n, on which the uninformative prior is applied.

3.1.1. Proposal Distributions for b_n and $\boldsymbol{g}_n$

For these three hidden variables, the proposal distributions are the transitional distributions in Eq. (9):

$$q(b_n|b_{n-1}^{(i)}) = p(b_n|b_{n-1}^{(i)}), \quad q(\boldsymbol{g}_n|\boldsymbol{g}_{n-1}^{(i)}) = p(\boldsymbol{g}_n|\boldsymbol{g}_{n-1}^{(i)}) \tag{17}$$

3.1.2. Proposal Distribution for $\boldsymbol{h}_n$

For $\boldsymbol{h}_n$, the proposal does not depend on previous frames, but on $\boldsymbol{g}_n$

$$q(\boldsymbol{h}_n|\boldsymbol{g}_n^{(i)}) = \mathcal{N}(\hat{\boldsymbol{h}}_n^{(i)}, s_h^2 I) \tag{18}$$

where s_h^2 is a hyper-parameter and $\hat{\boldsymbol{h}}_n^{(i)}$ is an estimate of $\boldsymbol{h}_n$ for particle i, using cepstral analysis. Specifically, denote the signal cepstrum by $c_n[k]$. In voiced case, voiced energy is assumed to dominate non-voiced energy. From Eqs. (2) and (4), the filter of voiced energy is $G_n(\omega)H_n(\omega)$. According to cepstral theory [15]

$$c_n[k] \approx g_n[k] + h_n[k], \forall k \text{ at low quefrency} \tag{19}$$

Stacking the low quefrency part of $c_n[k]$ and $\hat{g}_n^{(i)}[k]$ into vectors $\boldsymbol{c}_n$ and $\hat{\boldsymbol{g}}_n^{(i)}$ respectively, we have

$$\hat{\boldsymbol{h}}_n^{(i)} = \boldsymbol{c}_n - \hat{\boldsymbol{g}}_n^{(i)} \tag{20}$$

where $\hat{\boldsymbol{g}}_n^{(i)}$ is the estimate of glottal complex cepstrum from $\boldsymbol{g}_n^{(i)}$.

In unvoiced case, the transfer function is $H(\omega)$. Therefore

$$\hat{\boldsymbol{h}}_n^{(i)} = \boldsymbol{c}_n \tag{21}$$

3.1.3. Proposal Distributions for $f_{0,n}$ and τ_n

The proposal distributions for $f_{0,n}$ and τ_n are Gaussian Mixture Models (GMMs):

$$q(f_{0,n}) = \sum_l \gamma_l \mathcal{N}(\hat{f}_{0,n}(l), s_f^2), \quad q(\tau_n) = \sum_l \delta_l \mathcal{N}(\hat{\tau}_n(l), s_\tau^2) \tag{22}$$

where s_f^2 and s_τ^2 are hyper-parameters, $\{\hat{f}_{0,n}(l)\}$ are arithmetic inverse of lags corresponding to peaks of autocorrelation function; $\hat{\tau}_n$ are peak times of short-time energy function. γ_l and δ_l are proportional to peak height.

3.2. Inner Layer Inference using Taylor Expansion Assisted M-CMC

The inner layer inference uses the Metropolis-Hastings (MH) algorithm in the MCMC framework, to sample from $p(\boldsymbol{I}_n|\boldsymbol{O}_n, \boldsymbol{x}_n)$. Hereafter, this target distribution is written as $p(\boldsymbol{I}|\boldsymbol{O}, \boldsymbol{x})$, by suppressing frame subscript n, since the sampling is within each frame.

In each iteration, given the current sample $\boldsymbol{I}^{(r)}$, a new sample $\boldsymbol{I}$ is drawn from a proposal distribution $q(\boldsymbol{I}|\boldsymbol{I}^{(r)}, \boldsymbol{O}, \boldsymbol{x})$, and is accepted with probability

$$\min\left\{1, \frac{p(\boldsymbol{I}|\boldsymbol{O}, \boldsymbol{x})q(\boldsymbol{I}^{(r)}|\boldsymbol{I}, \boldsymbol{O}, \boldsymbol{x})}{p(\boldsymbol{I}^{(r)}|\boldsymbol{O}, \boldsymbol{x})q(\boldsymbol{I}|\boldsymbol{I}^{(r)}, \boldsymbol{O}, \boldsymbol{x})}\right\} \tag{23}$$

To define the proposal distribution q, inspired by [17], we exploit 2nd-order Taylor expansion of the single-site conditional distribution

$$\ln p\left(\boldsymbol{I}(t)|\boldsymbol{I}^{(r)}(1:t-1), \boldsymbol{I}^{(r)}(t+1:T), \boldsymbol{O}, \boldsymbol{x}\right) \tag{24}$$

A sensible point around which to take the Taylor expansion is the prior conditional mean of $\boldsymbol{I}(t)$, which has closed form as follows:

$$\begin{aligned}\boldsymbol{m}(t) &\triangleq \mathbb{E}\left(\boldsymbol{I}(t)|\boldsymbol{I}^{(r)}(t-2:t-1), \boldsymbol{I}^{(r)}(t+1:t+2)\right) \\ &= \kappa\left(\boldsymbol{I}^{(r)}(t-2) + \boldsymbol{I}^{(r)}(t+2)\right) + \iota\left(\boldsymbol{I}^{(r)}(t-1) + \boldsymbol{I}^{(r)}(t+1)\right)\end{aligned} \tag{25}$$

where $\kappa = (1-\lambda)/A$, $\iota = \lambda^2/A$, $A = 2(1-\lambda+\lambda^2)$.

Here we state without further derivation that the single-site proposal distribution $q\left(\boldsymbol{I}(t)|\boldsymbol{I}^{(r)}(1:t-1), \boldsymbol{I}^{(r)}(t+1:T), \boldsymbol{O}, \boldsymbol{x}\right)$ obeys 2-d independent Gaussian. Define $\Sigma \triangleq \text{diag}(\sigma_{0:L-1}^2)$ according to Eq. (6). Then for dimension $a = 1, 2$, the Gaussian mean $\mu_{t,a}$ and variance $\lambda_{t,a}$ are as follows:

$$\begin{aligned}&\mu_{t,a} = \boldsymbol{m}_a(t) - \lambda_{t,a}\left(\boldsymbol{s}^{(r,\boldsymbol{m}(t))} - \boldsymbol{x}\right)^T \Sigma^{-1}\nabla_{t,a}^{(r)} \\ &\lambda_{t,a}^{-1} = A\Sigma_{I,a} + \left(\nabla_{t,a}^{(r)}\right)^T \Sigma^{-1}\nabla_{t,a}^{(r)} + \left(\boldsymbol{s}^{(r,\boldsymbol{m}(t))} - \boldsymbol{x}\right)^T \Sigma^{-1}\nabla_{t,a}^{2,(r)}\end{aligned} \tag{26}$$

To understand this, first note that the voiced part, $\boldsymbol{s}$ is a deterministic vector function of $\boldsymbol{O}$ and $\boldsymbol{I}$, which could be denoted by $\boldsymbol{s}(\boldsymbol{O}, \boldsymbol{I})$ with some abuse of notation. Then, $\nabla_{t,a}^{(r)}$ and $\nabla_{t,a}^{2,(r)}$ are defined as the gradient and Hessian's diagonal of

$$\boldsymbol{s}\left(\boldsymbol{O}, \boldsymbol{I}^{(r)}(1:t-1), \boldsymbol{m}(t), \boldsymbol{I}^{(r-1)}(t+1:T)\right) \triangleq \boldsymbol{s}^{(r,\boldsymbol{m}(t))} \tag{27}$$

with respect to dimension a of $\boldsymbol{m}(t)$, $a = 1, 2$.

Finally, we obtain the full proposal distribution

$$q(\boldsymbol{I}|\boldsymbol{I}^{(r)}, \boldsymbol{O}, \boldsymbol{x}) \triangleq \prod_t q\left(\boldsymbol{I}(t)|\boldsymbol{I}^{(r)}(1:t-1, t+1:T), \boldsymbol{O}, \boldsymbol{x}\right) \tag{28}$$

4. EXPERIMENT

4.1. Configurations

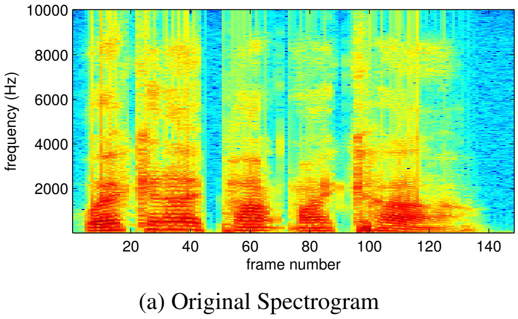
To evaluate the effectiveness of the new PAT3 model along with the newly developed inference algorithm, we adopt the Edinburgh Speech Corpus [18], which consists of a male speaker and a female speaker, each producing 50 sentences. The total length of 100 utterances is 331.6s, and the sample rate is 20KHz. The corpus has accurate glottal measurement by laryngograph, which provides valuable ground truth for GCI and pitch frequencies.

Each utterance is segmented into 30ms speech frames with 10ms frame shift. For each frame, the inferred values of hidden variables $\hat{\boldsymbol{O}}_n^{\text{MAP}}$ and $\hat{\boldsymbol{I}}_n$ are computed by the new PAT3.

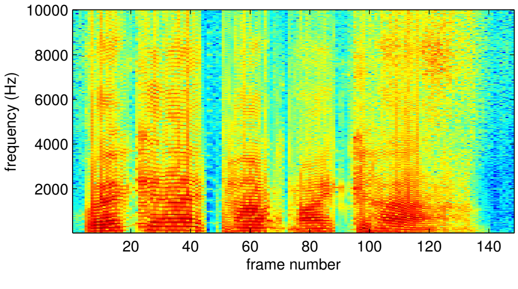
4.2. Signal Reconstruction and AM-FM Tracking

Reconstruction is the inferred $S_n(\omega)$ (voiced case) or $\tilde{b}_n U_n(\omega) H_n(\omega)$ (unvoiced case), which are both determined by the inferred $\boldsymbol{O}_n$ and $\boldsymbol{I}_n$. Fig. 1 compares the reconstructed spectrogram (top panel) and the original one (bottom panel) of utterance 1 of the male speaker. Both spectrograms are similar, particularly in low frequency, where the energy is dominated by voiced.

Fig. 2 plots the reconstructed voiced waveform from $S_n(\omega)$ of frame 107, utterance 1 of the male speaker. The black line is the original waveform, and the red line is the reconstructed waveform. The top panel shows the reconstructed waveform without AM-FM tracking, i.e. the inferred waveform with $\boldsymbol{O}_n$ set to its inferred value, but $\boldsymbol{I}_n$ (which contains of AM-FM related variables) set to constant.

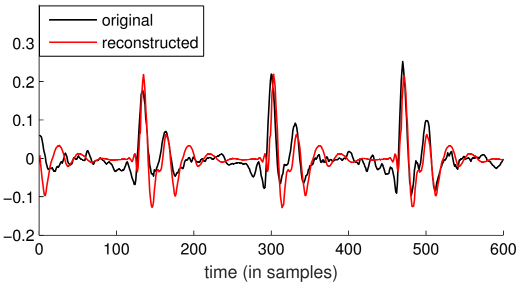


(a) Original Spectrogram

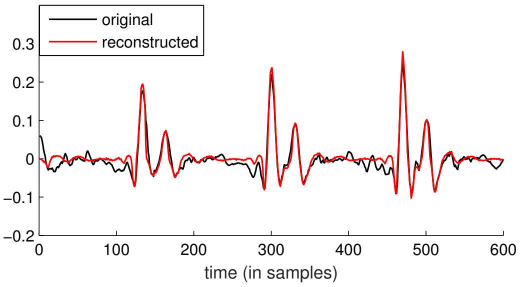


(b) Reconstructed Spectrogram

Fig. 1: Comparison of the reconstructed and original spectrogram.



(a) Without AM-FM Tracking



(b) With AM-FM Trakcing

Fig. 2: The reconstructed voiced waveform (red line) with and without AM-FM tracking, compared with the original waveform (black).

The bottom panel shows the result with AM-FM tracking, i.e. the inferred waveform with $\boldsymbol{O}_n$ and $\boldsymbol{I}_n$ both set to their inferred values. As can be seen, the AM-FM effect in the original waveform is significant - it can capture constant increase in amplitude and a slight decrease in F0.

Table 1 shows the averaged SNR of the reconstruction using PAT3 compared to that using STRAIGHT [5]. PAT3 exceeds STRAIGHT by almost 10dB. This is because STRAIGHT is a model for power spectrum only, not considering the phase, whereas PAT3 models the speech waveform itself.

PAT3	STRAIGHT
8.79	-2.57

Table 1: SNR of speech reconstruction

	PAT3	GetF0
GPE (%)	2.10	2.07
RMS (Hz)	5.052	5.780

Table 2: Pitch tracking results on Edinburgh dataset

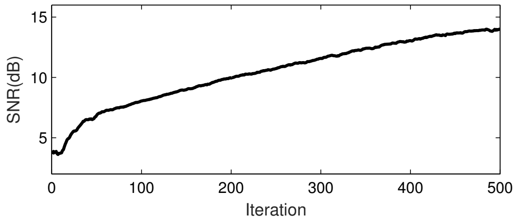


Fig. 3: SNR of reconstruction as a function of MCMC iterations.

Combining these results we can see that PAT3 can accurately reconstruct not only the amplitude, but also the phase with good accuracy. Phase modeling has been notoriously difficult in speech processing, yet important in applications such as speech denoising and glottal estimation.

4.3. Pitch Tracking

Pitch tracking by PAT3 is essentially the inference of $f_{0,n}$. Since a U/V decision scheme for PAT3 has yet to be developed, we extract pitch on labeled voiced segments only, and compare against a pitch-tracking benchmark, GetF0 [19]. Both algorithms are run over the complete Edinburgh dataset. For fair comparison, we compare the pitch tracking results of all the voiced frames that are also correctly classified as voiced by GetF0, in terms of the following 2 criteria:
Gross Pitch Error (GPE): The percentage of frames whose pitch estimates deviate from ground truth by more than 20%.
Root Mean Squared Error (RMS): The averaged mean squared error in Hz over the frames free of GPE.

Table 2 shows the results. As can be seen, PAT3 has comparable GPE level to GetF0, but much smaller RMS, which means PAT3 inference is more accurate.

4.4. Performance of the Sampling Inference

Fig. 3 plots the SNR as a function of inner MCMC iterations, which implies that it mixes well within 500 iterations. For efficiency in practice, we set 60 particles for outer layer APF, and 5 inner MCMC iterations for each particle. Then the best particle is perfected with another 100 iterations. The proposed inference algorithm, working with the 1200-dimensional AM-FM hidden variables $\boldsymbol{I}_n$ in each frame, in addition to the 55-dimensional $\boldsymbol{O}_n$, produces satisfactory performance, as shown in all the above results.

5. CONCLUSION

This paper introduces an AR process to model AM-FM of the voice source. The resulting model is complex, but its parameters can be inferred using an APF outer loop and a Taylor expansion assisted MCMC inner loop. Results demonstrate reconstruction of both signal magnitude and phase from probabilistically inferred generative model parameters with a reconstruction SNR of 8.79dB.

6. REFERENCES

[1] H. Kameoka, N. Ono, and S. Sagayama, “Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1507–1516, 2010.

[2] J. P. Cabral, “Hmm-based speech synthesis using an acoustic glottal source model,” Ph.D. dissertation, School of Informatics, The University of Edinburgh, 2011.

[3] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, “Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis,” *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2013.

[4] R. Turner and M. Sahani, “Probabilistic amplitude and frequency demodulation,” in *Advances in Neural Information Processing Systems*, 2011, pp. 981–989.

[5] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, “Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 3933–3936.

[6] Z. Ou and Y. Zhang, “Probabilistic acoustic tube: a probabilistic generative model of speech for speech analysis/synthesis,” in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012, pp. 841–849.

[7] Y. Zhang, Z. Ou, and M. Hasegawa-Johnson, “Improvement of probabilistic acoustic tube model for speech decomposition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7929–7933.

[8] ——, “Incorporating am-fm effect in voiced speech for probabilistic acoustic tube model,” in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.

[9] Y. Qi, T. P. Minka, and R. W. Picara, “Bayesian spectrum estimation of unevenly sampled nonstationary data,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002, pp. II–1473.

[10] J. Carpenter, P. Clifford, and P. Fearnhead, “Improved particle filter for nonlinear problems,” *IEE Proceedings-Radar, Sonar and Navigation*, vol. 146, no. 1, pp. 2–7, 1999.

[11] M. K. Pitt and N. Shephard, “Filtering via simulation: Auxiliary particle filters,” *Journal of the American statistical association*, vol. 94, no. 446, pp. 590–599, 1999.

[12] J. S. Liu, *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[13] P. Clark and L. E. Atlas, “Time-frequency coherent modulation filtering of nonstationary signals,” *Signal Processing, IEEE Transactions on*, vol. 57, no. 11, pp. 4323–4332, 2009.

[14] Y. Pantazis, O. Rosec, and Y. Stylianou, “Adaptive am–fm signal decomposition with application to speech analysis,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 290–300, 2011.

[15] T. Quariteri, *Discrete-time Speech Signal Processing*. Prentice-Hall, 2002, p. 64.

[16] S. Godsill, A. Doucet, and M. West, “Maximum a posteriori sequence estimation using monte carlo particle filters,” *Annals of the Institute of Statistical Mathematics*, vol. 53, no. 1, pp. 82–96, 2001.

[17] N. Shephard and M. K. Pitt, “Likelihood analysis of non-gaussian measurement time series,” *Biometrika*, vol. 84, no. 3, pp. 653–667, 1997.

[18] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, “Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching.” in *Proc. Eurospeech*. International Speech Communication Association, 1993.

[19] D. Talkin, “Robust algorithm for pitch tracking,” *Speech coding and synthesis*, pp. 497–518, 1995.